Appendix A: Full Dataset Construction and Feature Definitions

A.1 Transcript Collection and Normalization

CourtShadow uses 42 publicly available criminal trial transcripts sourced from state and federal courts, archival repositories, and journalist-compiled public collections. Because the documents vary substantially in formatting, a normalization pipeline was applied:

1. **Whitespace normalization:** Consecutive spaces replaced with a single space; stray indentation removed.
2. **Linebreak smoothing:** Page headers, footers, and transcription artifacts removed via regular-expression filtering.
3. **Speaker-tag alignment:** The system identifies lines of the form `THE COURT:`, `MR. SMITH:`, `DEFENDANT:`, etc. to recover speaker turns.
4. **Segment construction:** each transcript is divided into short, linguistically meaningful segments using:
   - Speaker-change boundaries,
   - Length heuristics (target: 6–70 tokens),
   - Punctuation cues (“.”, “?”, “!”).

   Segments with fewer than 6 tokens are merged with adjacent segments when linguistically appropriate.
5. **Speaker-role metadata:** Each segment is labeled as belonging to: judge, prosecutor, defense attorney, witness, defendant, or unknown.

In total, the processed dataset includes **1,915 segments** across **42 cases**.

A.2 Feature Engineering Overview

Each segment is represented as a 38-dimensional feature vector constructed from four interpretable families:

1. Structure (3 features)
2. Discursive Framing (10 features)

3. Pronouns & Voice (4 features)

4. Topic Indicators (21 features)

All features are computed from the raw text without syntactic parsers or pretrained models to preserve interpretability.

A.3 Structure Features (3)

- **chunk_tokens** — Approximate token count computed by splitting on whitespace.
- **chunk_sentences** — Count of end-of-sentence punctuation: ".", "!", "?".
- **chunk_question_marks** — Count of "?" as an approximation of adversarial or interrogative pressure.

These features quantify structural density and questioning intensity, grounding the model in basic discourse structure.

A.4 Discursive Framing Features (10)

These capture stance, evaluation, and interpersonal framing, drawing from sociolinguistic work on courtroom discourse (Conley & O'Barr 1990; Gibbons 2003; Maynard 1984).

- **politeness_count** — Instances of: {please, thank you, Your Honor, sir, ma'am}.
- **politeness_rate** — Politeness per token.
- **harsh_count** — Instances of: {dangerous, violent, criminal, liar, threat}.
- **harsh_rate** — Harsh terms per token.
- **mitigation_count** — {circumstances, complicated, unfortunate, remorse, treatment, background}.
- **certainty_strong** — {clearly, obviously, beyond doubt, without question}.
- **certainty_weak** — {maybe, possibly, might, perhaps}.
- **evaluation_positive** — {cooperative, respectful, honest, compliant}.

- **evaluation_negative** — {aggressive, noncompliant, evasive}.
- **sentiment_imbalance** — (positive – negative) surrogate measured with dictionary lookup.

A.5 Pronoun & Voice Features (4)

Based on prior work on institutional stance and courtroom power dynamics:

- **first_person_total** — Count of {I, me, my, we, us, our}.
- **first_person_rate** — First-person / tokens.
- **second_person_total** — Count of {you, your, yours}.
- **second_person_rate** — Second-person / tokens.

These encode who is narrating ("I") and who is addressed ("you"), which differ systematically across defendant groups.

A.6 Topic Indicator Features (21)

Topic indicators are binary flags that fire when a segment contains keywords drawn from five criminal-legal categories. These capture procedural context and case-type differences without relying on any external models.

Violence & Weapons (6)

- Guns, firearm, gunshot, knife, stabbing, assault, murder, manslaughter

Police Interaction (6)

- Police, officer, cop, arrest, comply, resist, flee, detained

Financial Crime (4)

- Fraud, scheme, bank, account, investment, wire transfer

Drugs (3)

- Drugs, narcotics, cocaine, heroin, meth

Harm & Injury (2)

- Victim, injury, injuries, hospital

A.7 Feature Scaling

All continuous features (e.g., counts, rates, length) are standardized using:

$$x' = \frac{x - \mu}{\sigma},$$

where μ and σ are computed from the **training** set only. Binary topic indicators are left unscaled.

A.8 Segment-Level to Case-Level Mapping

Each segment is tagged with:

- Segment text,
- Speaker role,
- Case ID,
- Inherited group label.

After model prediction, segment-level probabilities are aggregated via:

$$\bar{p}_{case} = \frac{1}{m} \sum_{j=1}^{m} p_j.$$

This produces the **Linguistic Environment Score** (LES), which smooths local variation and highlights institutional patterns across the transcript.

A.9 Data Quality Considerations

- Transcripts vary in completeness; some lack opening statements or cross-examination segments.
- Formatting irregularities (OCR noise, double-spacing, missing speaker tags) are corrected heuristically.
- Case-type heterogeneity (violent crimes, financial crimes, drug cases) is partially controlled by topic indicators.

- No transcripts include explicit race labels; group assignments come from public case metadata.

These constraints motivate the project's emphasis on interpretability and cautious inference.